Comparing Wikipedia and German Wordnet by Evaluating Semantic Relatedness on Multiple Datasets

Torsten Zesch and **Iryna Gurevych** and **Max Mühlhäuser**
Ubiquitous Knowledge Processing Group, Telecooperation Division
Darmstadt University of Technology, D-64289 Darmstadt, Germany
{zesch,gurevych,max} (at) tk.informatik.tu-darmstadt.de

Abstract

We evaluate semantic relatedness measures on different German datasets showing that their performance depends on: (i) the definition of relatedness that was underlying the construction of the evaluation dataset, and (ii) the knowledge source used for computing semantic relatedness. We analyze how the underlying knowledge source influences the performance of a measure. Finally, we investigate the combination of wordnets and Wikipedia to improve the performance of semantic relatedness measures.

1 Introduction

Semantic similarity (**SS**) is typically defined via the lexical relations of synonymy ($automobile$ – car) and hypernymy ($vehicle$ – car), while **semantic relatedness** (**SR**) is defined to cover any kind of lexical or functional association that may exist between two words. Many NLP applications, like sense tagging or spelling correction, require knowledge about semantic relatedness rather than just similarity (Budanitsky and Hirst, 2006). For these tasks, it is not necessary to know the exact type of semantic relation between two words, but rather if they are closely semantically related or not. This is also true for the work presented herein, which is part of a project on electronic career guidance. In this domain, it is important to conclude that the words “baker” and “bagel” are closely related, while the exact type of a semantic relation does not need to be determined.

As we work on German documents, we evaluate a number of SR measures on different German datasets. We show that the performance of measures strongly depends on the underlying knowledge source. While WordNet (Fellbaum, 1998) models SR, wordnets for other languages, such as the German wordnet GermaNet (Kunze, 2004), contain only few links expressing SR. Thus, they are not well suited for estimating SR.

Therefore, we apply the Wikipedia category graph as a knowledge source for SR measures. We show that Wikipedia based SR measures yield better correlation with human judgments on SR datasets than GermaNet measures. However, using Wikipedia also leads to a performance drop on SS datasets, as knowledge about classical taxonomic relations is not explicitly modeled. Therefore, we combine GermaNet with Wikipedia, and yield substantial improvements over measures operating on a single knowledge source.

2 Datasets

Several German datasets for evaluation of SS or SR have been created so far (see Table 1). Gurevych (2005) conducted experiments with a German translation of an English dataset (Rubenstein and Goodenough, 1965), but argued that the dataset (**Gur65**) is too small (it contains only 65 noun pairs), and does not model SR. Thus, she created a German dataset containing 350 word pairs (**Gur350**) containing nouns, verbs and adjectives that are connected by classical and non-classical relations (Morris and Hirst, 2004). However, the dataset is biased towards strong classical relations, as word pairs were manually selected. Thus, Zesch and Gurevych (2006) semi-automatically created word pairs from domain-specific corpora. The resulting **ZG222** dataset contains 222 word pairs that are connected by all kinds of lexical semantic relations. Hence, it is particularly suited for analyzing the capability of a measure to estimate SR.

Proceedings of NAACL HLT 2007, Companion Volume, pages 205–208,
Rochester, NY, April 2007.

								CORRELATION r	
DATASET	YEAR	LANGUAGE	# PAIRS	POS	TYPE	SCORES	# SUBJECTS	INTER	INTRA
Gur65	2005	German	65	N	SS	discrete {0,1,2,3,4}	24	.810	-
Gur350	2006	German	350	N, V, A	SR	discrete {0,1,2,3,4}	8	.690	-
ZG222	2006	German	222	N, V, A	SR	discrete {0,1,2,3,4}	21	.490	.647

Table 1: Comparison of datasets used for evaluating semantic relatedness.

3 Semantic Relatedness Measures

Semantic wordnet based measures Lesk (1986) introduced a measure (***Les***) based on the number of word overlaps in the textual definitions (or glosses) of two terms, where higher overlap means higher similarity. As GermaNet does not contain glosses, this measure cannot be employed. Gurevych (2005) proposed an alternative algorithm (***PG***) generating surrogate glosses by using a concept's relations within the hierarchy. Following the description in Budanitsky and Hirst (2006), we further define several measures using the taxonomy structure.

$$
\begin{aligned}
sim_{PL} &= l(c_1, c_2) \\
sim_{LC} &= -\log \frac{l(c_1, c_2)}{2 \times depth} \\
sim_{Res} &= IC(c_i) = -\log(p(lcs(c_1, c_2))) \\
dist_{JC} &= IC(c_1) + IC(c_2) - 2IC(lcs(c_1, c_2)) \\
sim_{Lin} &= 2 \times \frac{IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)}
\end{aligned}
$$

PL is the taxonomic path length $l(c_1, c_2)$ between two concepts $c1$ and $c2$. ***LC*** normalizes the path length with the depth of the taxonomy. ***Res*** computes SS as the information content (IC) of the lowest common subsumer (lcs) of two concepts, while ***JC*** combines path based and IC features.[1] ***Lin*** is derived from information theory.

Wikipedia based measures For computing the SR of two words w_1 and w_2 using Wikipedia, we first retrieve the articles or disambiguation pages with titles that equal w_1 and w_2 (see Figure 1). If we hit a redirect page, we retrieve the corresponding article or disambiguation page instead. In case of an article, we insert it into the candidate article set (A_1 for $w1$, A_2 for w_2). In case of a disambiguation page, the page contains links to all encoded word senses, but it may also contain other links. Therefore, we only consider links conforming to the pattern $\langle$Title_(DisambiguationText)$\rangle$[2] (e.g. “Train_(roller coaster)”). Following all such links gives the candidate article set. If no disambiguation links are found, we take the first link on the page, as most important links tend to come first. We add the corresponding articles to the candidate set. We form pairs from each candidate article $a_i \in A_1$ and each article $a_j \in A_2$. We then compute $SR(a_i, a_j)$ for each pair. The output of the algorithm is the maximum SR value $\max_{a_i \in A_1, a_j \in A_2}(SR(a_i, a_j))$.[3]

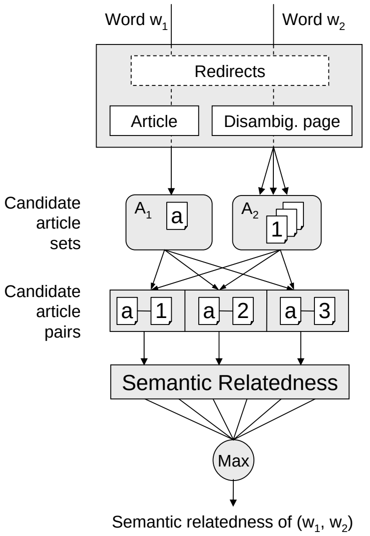

Figure 1: Steps for computing SR using Wikipedia.

As most SR measures have been developed for taxonomic wordnets, porting them to Wikipedia requires some modifications (see Figure 2). Text overlap measures can be computed based on the article text, while path based measures operate on the category graph. We compute the overlap between article

[1] Note that JC returns a distance value instead of a similarity value resulting in negative correlation with human judgments.

[2] ‘_(DisambiguationText)’ is optional.

[3] Different from our approach, Strube and Ponzetto (2006) use a disambiguation strategy that returns only a single candidate article pair. This unnecessarily limits a measure's potential to consider SR between all candidate article pairs. They also limit the search for a lcs to a manually specified threshold of 4.

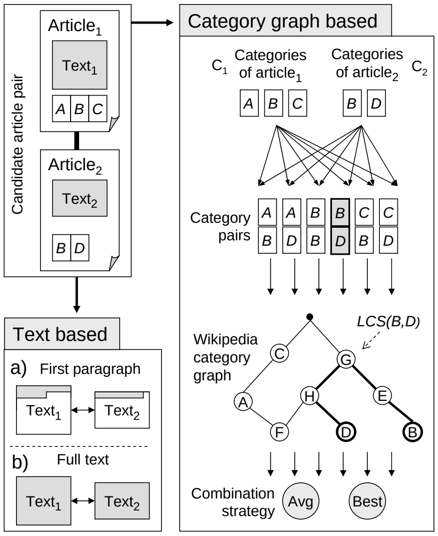

Figure 2: SR measures adapted on Wikipedia.

texts based on (i) the first paragraph, as it usually contains a short gloss, and (ii) the full article text. As Wikipedia articles do not form a taxonomy, path based measures have to be adapted to the Wikipedia category graph (see the right part of Figure 2). We define C_1 and C_2 as the set of categories assigned to article a_i and a_j, respectively. We compute the SR value for each category pair (c_k, c_l) with $c_k \in C_1$ and $c_l \in C_2$. We use two different strategies to combine the resulting SR values: First, we choose the best value among all pairs (c_k, c_l), i.e., the minimum for path based, and the maximum for information content based measures. As a second strategy, we average over all category pairs.

4 Experiments & Results

Table 2 gives an overview of our experimental results on three German datasets. Best values for each dataset and knowledge source are in bold. We use the PG measure in optimal configuration as reported by Gurevych (2005). For the Les measure, we give the results for considering: (i) only the first paragraph (+First) and (ii) the full text (+Full). For the path length based measures, we give the values for averaging over all category pairs (+Avg), or taking the best SR value computed among the pairs (+Best). For each dataset, we report Pearson's correlation r with human judgments on pairs that are found in **both resources** (**Both**). Otherwise, the results would not be comparable. We additionally use a subset containing only **noun-noun pairs** (**Both NN**). This comparison is fairer, because article titles in Wikipedia are usually nouns. Table 2 also gives the inter annotator agreement for each subset. It constitutes an upper bound of a measure's performance.

Our results on Gur65 using GermaNet are very close to those published by Gurevych (2005), ranging from 0.69–0.75. For Gur350, the performance drops to 0.38–0.50, due to the lower upper bound, and because GermaNet does not model SR well. These findings are endorsed by an even more significant performance drop on ZG222. The measures based on Wikipedia behave less uniformly. Les yields acceptable results on Gur350, but is generally not among the best performing measures. LC +Avg yields the best performance on Gur65, but is outperformed on the other datasets by PL +Best, which performs equally good for all datasets.

If we compare GermaNet based and Wikipedia based measures, we find that the knowledge source has a major influence on performance. When evaluated on Gur65, that contains pairs connected by SS, GermaNet based measures perform near the upper bound and outperform Wikipedia based measures by a wide margin. On Gur350 containing a mix of SS and SR pairs, most measures perform comparably. Finally, on ZG222, that contains pairs connected by SR, the best Wikipedia based measure outperforms all GermaNet based measures.

The impressive performance of PL on the SR datasets cannot be explained with the structural properties of the category graph (Zesch and Gurevych, 2007). Semantically related terms, that would not be closely related in a taxonomic wordnet structure, are very likely to be categorized under the same Wikipedia category, resulting in short path lengths leading to high SR. These findings are contrary to that of (Strube and Ponzetto, 2006), where LC outperformed path length. They limited the search depth using a manually defined threshold, and did not compute SR between all candidate article pairs.

Our results show that judgments on the performance of a measure must always be made with respect to the task at hand: computing SS or SR. Depending on this decision, we can choose the best underlying knowledge source. GermaNet is better for

		Gur65	Gur350		ZG222	
		Both NN	Both	Both NN	Both	Both NN
# Word Pairs		53	116	91	55	45
Inter Annotator Agreement		0.80	0.64	0.63	0.44	0.43
GermaNet	PG	0.69	0.38	0.42	**0.23**	0.21
	JC	**-0.75**	**-0.52**	-0.48	-0.19	**-0.25**
	Lin	0.73	0.50	**0.50**	0.08	-0.12
	Res	0.71	0.42	0.42	0.10	0.13
Wikipedia	Les +First	0.16	0.36	0.32	0.01	0.11
	Les +Full	0.19	0.34	0.37	0.13	0.17
	PL +Avg	-0.32	-0.34	-0.46	-0.36	-0.43
	PL +Best	-0.35	**-0.42**	**-0.53**	**-0.43**	**-0.49**
	LC +Avg	**0.37**	0.25	0.34	0.30	0.30
	LC +Best	0.21	0.12	0.21	0.15	0.12
Combination	Linear	**0.77**	**0.59**	**0.60**	0.38	0.43
	POS	-	0.55	-	**0.48**	-

Table 2: Correlation r of human judgments with SR measures on different datasets.

estimating SS, while Wikipedia should be used to estimate SR. Therefore, a measure based on a single knowledge source is unlikely to perform well in all settings. We computed a linear combination of the best measure from GermaNet and from Wikipedia. Results for this experiment are labeled *Linear* in Table 2. *POS* is an alternative combination strategy, where Wikipedia is only used for noun-noun pairs. GermaNet is used for all other part-of-speech (POS) combinations. For most datasets, we find a combination strategy that outperforms all single measures.

5 Conclusion

We have shown that in deciding for a specific measure and knowledge source it is important to consider (i) whether the task at hand requires SS or SR, and (ii) which POS are involved. We pointed out that the underlying knowledge source has a major influence on these points. GermaNet is better used for SS, and contains nouns, verbs, and adjectives, while Wikipedia is better used for SR between nouns. Thus, GermaNet and Wikipedia can be regarded as complementary. We have shown that combining them significantly improves the performance of SR measures up to the level of human performance.

Future research should focus on improving the strategies for combining complementary knowledge sources. We also need to evaluate a wider range of measures to validate our findings. As the simple PL measure performs remarkably well, we should also consider computing SR based on the Wikipedia article graph instead of the category graph.

Acknowledgments

This work was supported by the German Research Foundation under grant "Semantic Information Retrieval from Texts in the Example Domain Electronic Career Guidance", GU 798/1-2.

References

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based Measures of Semantic Distance. *Computational Linguistics*, 32(1).

Christiane Fellbaum. 1998. *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Iryna Gurevych. 2005. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proc. of IJCNLP*, pages 767–778.

Claudia Kunze, 2004. *Lexikalisch-semantische Wortnetze*, chapter Computerlinguistik und Sprachtechnologie, pages 423–431. Spektrum Akademischer Verlag.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone. In *Proc. of the 5th Annual International Conference on Systems Documentation*, pages 24–26.

Jane Morris and Graeme Hirst. 2004. Non-Classical Lexical Semantic Relations. In *Proc. of the Workshop on Computational Lexical Semantics, NAACL-HTL.*

Herbert Rubenstein and John B. Goodenough. 1965. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.

Michael Strube and Simone Paolo Ponzetto. 2006. WikiRelate! Computing Semantic Relatedness Using Wikipedia. In *Proc. of AAAI*, pages 1219–1224.

Torsten Zesch and Iryna Gurevych. 2006. Automatically Creating Datasets for Measures of Semantic Relatedness. In *Proc. of the Workshop on Linguistic Distances, ACL*, pages 16–24.

T. Zesch and I. Gurevych. 2007. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proc. of the TextGraphs-2 Workshop, NAACL-HLT*, (to appear).